The Ruth H. Hooker Research Library

and Technical Information Center

OPTICAL STORAGE AND RETRIEVAL OF LIBRARY MATERIAL

Doris R. Folen
Kestrel Associates
Laurie E. Stackpole, Chief Librarian

- **ABSTRACT**

The Ruth H. Hooker Research Library and Technical Information Center of the Naval Research Laboratory has installed an optical disk system consisting of a Sony autochanger, Sun minicomputer, Sun workstations, TDC scanners, printers, personal computers and various other peripherals. The system stores large portions of the Library's collection on 12-inch optical disks and can be expanded to allow retrieval over the campus network by the scientists of the Naval Research Laboratory. The first segment of the collection to be processed is a technical report collection consisting of 140,000 reports averaging 55 pages each. A third of this collection has currently been scanned to disk and is available for retrieval and on-demand printing by the Library patron.

- **INTRODUCTION**

This paper describes the development of an optical disk storage capability at the Ruth H. Hooker Research Library and Technical Information Center of the Naval Research Laboratory (NRL). Optical storage was selected for the preservation and maximum protection of the Library's immensely valuable collection of technical reports. This collection represents results of research in the area of physics and engineering since the beginning of the Second Wold War and much of this information is unavailable from any other source.

Optical disk technology also provides a viable, sensible solution to many of the serious recurrent problems that plague librarians. The most striking advantage of optical storage is that it provides a permanent solution to the space problem; no matter how you treat them, index them or catalog them, paper products, film products, etc. have to be somewhere taking up space. Space costs money. Omitting the costs associated with filing, retrieving and refiling reports, the savings to the NRL Laboratory in overhead will be over $100,000 a year when the entire unclassified report collection has been put on disk. Space requirements will drop from 3600 square feet to only 144 square feet. Optical storage saves money not only in charges associated with space but eliminates all future filing, retrieval and the refiling associated with physical storage. A paper copy of an item can be printed at the touch of a key and never has to be refiled since the archived copy remains on the disk. This technology presents a dream scenario in which a librarian can sit in a pleasant environment, identify a paper, report or picture and retrieve the item in seconds without leaving the area or the patron.

While the costs of equipment and conversion mandate careful consideration and planning for the implementation of optical storage and retrieval these costs are not so great that smaller libraries are prohibited from taking advantage of this technology. Small turn-key units can be purchased at reasonable costs and, if libraries collaborate in the scanning, costs can be absorbed in the budgets of even the smaller establishments.

Optical disk technology exists today and is being applied both in the Government and Commercial sector. For example: the Internal Revenue Service, which must deal with the phenomenal problem of storing income tax returns both from the private and industrial sector, has several prototype optical disk storage and retrieval systems currently operational and several others in the planning stage[1]. The giant insurance company, United

Services Automobile Association in San Antonio, Texas has produced a practically paperless operation using optical disk technology[2]; smaller insurance agencies are opting for this technology as well. The U.S. Navy's "paperless ship project" launched by Admiral Metcalf in 1987[3] is an undertaking which includes converting vast amounts of documents to optical storage. The list of those who are applying this technology is a long one, and users and potential users may be found wherever there are large quantities of information to be stored. Libraries are obvious beneficiaries as they are by their very nature places where large quantities of information is stored in every kind of format: books, journals, maps, charts, film, and computer databases. All are candidates for optical storage and retrieval.

The National Archives[4] and the Library of Congress[5] have both been working with prototypes for optical storage for over five years and have proven the importance of this technology for preserving their immensely valuable, but sometimes old and brittle documents. Libraries, including those at Cornell University[6] and Carnegie Mellon University[7] have designed optical storage and retrieval systems and currently are engaged in projects to exploit the capabilities of these systems. The NRL Library has had optical storage and retrieval capabilities fully operational for over three years. This optical disk system is one of the first of its kind and is still the largest, in terms of volumes of material stored, currently operational in a research library.

- **PLANNING FOR AN OPTICAL DISK SYSTEM**

Planning for an optical disk system is no easier now than it was four years ago although the problems are different. The major concern then was collecting information and advice since there were comparatively few people who were authorities in this field and very few facilities that had any experience with optical storage as the technology was so new at that time. Present day problems are due largely to the proliferation of new equipment and the burgeoning of technology.

Optical disk systems are expensive. They require right from the beginning a commitment of time and money. The first year cost of the NRL Library's system was just over $500,000. This included development of a prototype system, installation, maintenance and the scanning to optical disk of 15,000 reports. Additionally, upgrades to improve scanning, printing and OCR recognition have added another $500,000 to the cost. Also, many uncounted hours of staff time went into the original planning.

An optical disk system is a project that once started cannot easily be abandoned. It is going to an integral part of the library for a long time and plans for the future have to be considered. Thought must be given to who is going to use the system, who is going to run the system, and who is going to be responsible for its maintenance and future development. Level of staffing must be considered. Professional librarians and support technicians must be assigned to assist patrons and to maintain the system.

Funds for staffing must be budgeted depending on the size and planned use of the system. Large in-house systems require a supervisory librarian with knowledge of large computer systems and a staff that includes, at a minimum, reference librarians trained to use the system, a computer systems analyst or other system administrator, and library technicians or clerks to assist with the report conversion. A small library with minimum turn-key equipment may be able to absorb the entire operation with its existing staff.

Consideration should be given to the time requirements of an optical system. It will take time to plan the system, buy it, install it, run it, maintain it, and scan documents into it. Staff and patrons must be trained to use it, which will also take time.

A study should be made of why an optical disk system is needed: what problems is it going to solve, how is it going to be used, how will it change the way staff and patrons do their work. Decisions have to be made on design considerations and trade-offs. For example: is speed of retrieval important enough to justify increased system costs? or what is he best compromise between image quality and time and money available for the scanning process?

With currently available optical equipment, images from almost any media can be processed and sent to the optical disk. Equipment exists for automatically scanning film, microfiche, maps, charts, photographs, and

aperture cards. Therefore, decisions have to be made as to what is going to be converted to optical images. For example, the NRL Library has one million reports in microfiche format and they take up quite a bit of space and they are a bother to file. However, they are much more space efficient than paper products, so the decision was made to process the paper products. Vendors must be considered and selected. Once one or more vendors can be singled out as likely candidates discussions with them will identify the equipment most suitable incorporation into an appropriate system. Usually at least three or four configurations slowly emerge for consideration.

Money must be budgeted. This is a "chicken and egg" proposition. The amount of money available will determine the design. On the other hand, a study of what a system could accomplish for the organization might justify a higher budget. At this time, at least $250,000 would be needed for the purchase of the equipment to begin a large operation, this would include an autochanger for 12-inch disks, minicomputer, scanner, printer, supporting PC's and workstations. On the other end of the scale a small operation might cost only $50,000 which could provide a PC-run operation involving a single disk drive and 5.25-inch disks. System sizes can be looked at in terms of numbers of pages to be stored. A 12-inch optical disk can hold up to 130,000 pages while a 5.25-inch disk could store approximately 20,000 pages. If a collection consists of millions of pages to be stored, then a large system must be considered.

On-going costs in both large and smaller systems involve library staff to help patrons, purchase of equipment upgrades, equipment and software maintenance, and the cost of ongoing document conversion.

A good estimate of the money to be budgeted can only be arrived at after all aspects of an optical storage have been considered and in-house needs established. Costs of hardware and software are not something that a librarian/planner can determine alone as most equipment and/or systems are sold only through vendors. Extensive talks with vendors who are experts in the field of costs are needed to begin to solidify an estimate.

The future of the system must be considered. Evidence indicates that images can be safely stored on optical disks for up to a hundred years[8]. However, the technology is advancing very quickly. For example, images stored on a disk holding 3.2 gigabytes take up twice as many disks as images stored on recently available disks holding 6.55 gigabytes. Already the NRL Library has upgraded its system to accommodate the higher density disks. Multimedia storage, that is the recording of sound, video and graphics all on the same disk, is a reality and soon will be practical. Scanners and printers that handle color are available. Image scanners and printers that process sheets of paper at over 100 pages per minute are to be available at reasonable cost in the near future. Although not all of the technology that vendors and manufacturers promise for the future materializes, much of it does reach the market: provisions for upgrading an optical disk system to incorporate these advances should be part of the overall game plan.

- **THE LIBRARY'S OPTICAL DISK SYSTEM**

The NRL Library initially chose Online Computer Systems, Inc. as the OEM (Original Equipment Manufacturer). Online Computer Systems took the Library's choice of basic equipment and working with the library staff on the design, put together one of the first optical disk systems of major proportions to be used in a research library (see figure 1). Currently, Kestrel Associates, Inc. is providing technical staff with responsibility for the maintenance of equipment, report preparation and scanning, and the integration of new hardware and software into the system.

When the Library began its optical storage and retrieval project, there were very few vendors in this field. Now there are many well qualified companies who can provide systems tailored to any library's budget or need. One source of a listing of available vendors is The Handbook of Optical Memory Systems by C. Peter Waegemann[9] which has a chapter on Vendors and Consultants and is revised periodically. Proper selection of equipment is of immense importance. Just the cost alone necessitates a carefully considered decision. Once purchased, the major components become the platform for future developments. Replacement of major components are not easily made.

Many hours were spent at conventions and shows looking at equipment and talking to vendors before the equipment forming the nucleus of the NRL's optical disk system was selected. Dreamware and fantasyware

had to be identified as such.

Since the system was to be an in-house stand alone archive, standards were not a primary consideration in the selection of equipment. Standards for the 12-inch optical disk were in the talking stage at that time as they are still. The library worked around this problem by trying to project standards and by picking leaders in the field as vendors.

A Sony Writable Optical Disk Autochanger Model WDA-610 (see figure 2) and a Sun minicomputer and its work stations form the nucleus of the system. To be consistent with the market Sony now calls its autochanger a "jukebox" which it somewhat resembles in operation. It is designed specifically for use with Sony's 12-inch optical disks. The jukebox has a footprint of eight square feet and looks from the front remarkably like a two door refrigerator.

One 12-inch "write once read many" (WORM) optical disk can provide 6.55 gigabytes (GB) of digital data, enough to store the contents of 130,000 typewritten pages. With the ability to accommodate 50 disks of 6.55 gigabytes, one autochanger provides the equivalent storage space of up to 500 file cabinets. Up to four autochangers can be daisy chained to expand storage capacity to 1.312 terabytes of on-line data on a single SCSI interface. Through the synchronized use of two writable disk drives, the autochanger offers an average disk-to-disk access time of just 5.0 seconds.

Connected to the Sony autochanger by a SCSI interface is a Sun 3/280. It has a console, four work stations, a 892 MB hard disk, 16 MB ECC RAM, online Ethernet transceiver and a Ciprico SCSI controller. It also has a 9-track tape drive.

One of the in-house scanners is a TDC Docuscan DS-2600. It is capable of scanning pages of different sizes and thicknesses and can scan two sides of an 8 1/2" x 11" page in less than two seconds at 200 dpi. It is of compact design, having a footprint of 2' x 2' and weighs 100 lbs. This scanner is supported by an AST 386/33 microcomputer containing an Xionics XIP-B scanner card. Here the scanned images are compressed and stored on a 300 MB hard disk before being transferred to the Sun 3/280 where the images are processed and sent to the Sony autochanger.

In using the TDC DocuScan DS-2600 scanner, the operator slides the sheet to be scanned along the alignment guide. The transport system grips the sheet and carries it on a straight path past the scanning area and into the receiver tray. Thus paper jams are virtually eliminated. Scanning up to 2,000 pages an 8 hour work day is considered possible with this equipment.

A TDC DocuScan DS-4530 scanner has been incorporated into the system as the primary scanner. This scanner scans at 300 dpi, has sheet feeder, a monitor, and is capable of sustained scanning at a rate of 40 pages a minute. After enhancing the image electronically, this scanner compresses the image into industry-standard CCITT Group IV format. Using this scanner in conjunction with the TDC DS- 2600 the capability of scanning 16,000 pages in one 8 hour work day becomes a reality.

When a stored image is to be viewed it is retrieved by the Sun 3/60, decompressed, and either viewed at a work station or queued on a PC and printed out on an HP Laser-Jet printer. The equipment is networked internally on an Ethernet LAN, and could be connected to the campus-wide network, which is part of the Internet.

All bindings, staples, etc. that hold a report together must be removed before scanning. Documents can be scanned more readily if edges are fairly even, and if an automatic document feeder is used on the scanner it is almost imperative that edges be even. For "exploding" reports and cutting even edges a heavy duty paper cutter is necessary. The Library purchased a Challenge Model 20 (see figure 3). It has a height of 54 inches, a footprint of 3 x 4 feet, and weighs 530 pounds. It requires two hands on the control panel for operation to keep the operator out of harm's way. A Ferrups 18KVA Uninterruptible Power Supply was installed to provide power in the event of a general blackout. This system provides 14 minutes of full load power or 39 minutes half load which gives the entire computer operation time to take itself down in an emergency without any loss of data.

The equipment described above is all housed in the Documents Section of the Naval Research Library. The smaller units such as the work stations, printers and paper cutter are easily moved and are rearranged from time to time. The larger units such as the minicomputer, autochanger and emergency uninterruptible power supply are much more unwieldy and can be moved only with great care and difficulty. Environmental factors such as heat and cold are not a problem.

A comprehensive back-up routine has been devised consisting of daily tape back-ups of everything that has been scanned that day, disk-to-disk back-up when an optical platter has been filled, and a back-up of all software on Exobyte tape.

The Library staff involved with this optical disk system consists of a federal manager, who oversees the whole operation, two professional librarians who, as part of their general referral duties, help users identify needed reports, three contract computer technicians who do the report preparation, scanning, and routine maintenance, and one contract computer analyst who installs new hardware and software, takes care of hardware and software problems, and advises the supervisory librarian on proposed system enhancements.

- **THE CONVERSION PROCESS**

The Library's unrestricted report collection consists of roughly 140,000 reports averaging 55 pages per report. Before the conversion process began this collection occupied a space of 3600 sq. ft. A large percentage of the older reports were not in very good shape as time, insects, dirt, water and other calamities had taken their toll.

The reports that are scanned take a one way trip to the scanner. The quality of the images is consistently good enough to allow the destruction of the reports after they have been scanned. Nothing is gained if the reports are scanned and then returned to their place on the shelf to continue to take up space. There are, of course, some exceptions. Occasionally there is a report that has actual historical value and the original is considered worth preserving. Those few reports are scanned and then saved in an historical file.

Because this entire collection is to be scanned, no particular care has been taken with scanning order. Reports that have come back from circulation are put aside to scan because there is no sense filing them and then taking them off the shelf later to scan. New NRL originated reports are scanned as soon as the Library receives a copy. Reports are taken from the shelves in sequence but are not checked for misfiling as they are prepared for scanning because unlike full size paper reports which are filed on a shelf, images do not have to be placed on the disk in any particular order.

Time spent in preparation is well invested. Each report is reviewed as a candidate for scanning. Duplicates and reports without technical substance are weeded out. Reports that are oversized or of poor quality are put aside for later consideration.

A report is prepared for the scanner by:

1. Removing all blank and extraneous pages especially covers when possible. Usually all of the information on the cover is repeated on the title page. The back covers seldom have anything on them at all.

2. For in-house scanning the reports are then "exploded" by a quick cut by the paper cutter. This removes all binding, staples, etc. If necessary, additional trimming of the sides is done also.

3. Oversize pages and fold-outs are cut to 8 1/2" x 11" or less.

4. Loose color pictures are removed and filed.

It is important to establish a routine for the work flow if there are a lot of reports to be scanned (see figure 4). An occasional report to be scanned is one thing. When there are 140,000 of them to be reduced to images, efficiency becomes the watchword.

Use of the TDC DS-4530 scanner, is almost fully automatic. Pages are sent by a document feeder through

the scanner at a rate of 40 pages per minute. Bar coded accession numbers are read by the scanner and entered. The bar code also signals the first page of the next document. Quality of the scanned image is checked page by page on a preview monitor by the operator as the pages are scanned. A screen on an Austin 486/33 PC follows the progress of the document as it is scanned and the images compressed. The Austin PC also records scanning statistics.

When using the TDC DS-2600, the scanning operation is prompted by menu driven cues on the AST 386 terminal screen. The operator keys in the accession number, which is the only indexing done by the operator, and then follows a short series of instructions that appear on the screen as the report is fed page by page through the scanner. As a final step the scanned image is checked for quality. The report is then discarded.

Because it is the image of the item that is stored, graphs, diagrams, charts, black and white photographs and other items are scanned the same way as a printed page. Color pictures are set aside after scanning, tagged with their accession number and filed. They are then available if a library patron wishes to see them. Fortunately, because of the nature of the reports being scanned, there is not a large percentage of reports with colored pictures. The Library plans to scan these color pictures for retrieval with the rest of the original report when future technology produces viable and inexpensive equipment for scanning, retrieving and printing color images.

Both the TDC DS-2600 scanner and the new TDC DS-4530 scanner are in operation with the combined capability of scanning 16,000 pages in an eight hour work day. The costs incurred for in-house scanning using just the TDC DS-2600 was approximately $0.16 a page. The cost for scanning using both the TDC DS-2600 the new high speed TDC DS-4530 is estimated to be less than $0.10 a page. These costs includes report preparation, scanning, overhead, salary, and equipment amortization.

Because the Library is under a time constraint to accomplish the conversion to optical storage and retrieval, off-site scanning has been used as a means to accelerate the process. Now that the new high speed scanner is operational the project will be conducted entirely in-house.

In using off-base scanning facilities, reports are screened and prepared for scanning much the same as if they were going to be scanned in-house. Then they are shipped in boxes to Docucon Corporation in San Antonio, Texas. Docucon scans the reports at 300 dpi and records the images to nine track tape which is then sent to the NRL Library to be mounted on the Sun 3/280 for the images to be sent to disk.

Off-site scanning cost to the Library is figured at $0.21 a page. This cost includes report preparation, Docucon's cost per page, shipping, the Library's processing of the returned tapes, and the Library's overhead and equipment amortization.

- **RETRIEVAL**

The retrieval of a report from the optical disk system is initiated by typing the report's six digit accession number into a user friendly menu on a Sun work station (see figure 5). The report can then be viewed page by page on the screen or the viewer can skim through the report, back and forth, at the viewers own pace. All or part of the report can be printed out by a simple command. More than one report can be viewed on a screen and the same report can be viewed simultaneously at more than one work station at one time.

The ability to identify the report or reports to be retrieved has become a stumbling block for many system designers. Hundreds of thousands of pages of records placed on optical disk can rapidly submerge individual items in a ocean of data. Some designers solve this problem by superimposing an indexing system directly on the optical disk, but this takes a lot of time and disk space. Some designers superimpose ASCII characters that can be searched word by word directly on the disk, but this also takes a lot of disk space. Selecting an indexing system can take as much or more time as selecting the hardware for the Optical Disk System. An indexing program can be expensive and if the wrong one is chosen very costly indeed, and if the cataloging information is superimposed on the optical disk, it will be very difficult to change.

The NRL Library finessed this retrieval dilemma by opting to keep the retrieval system entirely separate

from the Optical Disk System.

The reasoning was:

1. An automatic retrieval system was already in place;

2. There was a need for the quickest and the most expedient means of getting the reports to optical disk and a separate retrieval system supported that goal;

3. A separate system is less expensive and easier to manage, input can be done at its own pace, and anything on the system can be changed at any time with little fuss;

4. At the beginning of the planning stage the optical disk system was thought of as simply an alternative way of storing reports;

5. The time saved by combining indexing and retrieval was not considered worth the additional cost and effort that it would have been necessary at the time the Library started this project. These reports do not have to be retrieved in tenths of a second. Retrieval of a report in less than half a minute is considered satisfactory. Compared with the time necessary to retrieve a dusty report from a real and usually dustier shelf, a half minute seems instantaneous.

Since August of 1987, the Library has had in place the Cuadra STAR retrieval system. This system provides a very large number of fields so that a report can be indexed in every conceivable way that someone might think of to identify it; i.e. accession number, title, author, subject, contract, words in an abstract, size, etc. Searching may be done by using any field separately or combined. The fields may be searched in full or by individual words, by subfields or by masking. Boolean searching of combinations of fields or search results is also possible. The results of all searches can be displayed or printed out by any or all of the fields used in the indexing in any order or in any combination.

The flexibility of the system allows for the addition of a field or subfield if it is discovered that there is a need for one that hadn't been thought of before. Global changes allow for sweeping changes in indexing if that should happen to be deemed necessary. Authority lists and lookup tables are used for controlling subject and descriptive cataloging terms, but fields are also available for the cataloger to use for flights of fancy if desired. The Library enjoys the reputation of being able to identify and provide any report that the patron needs.

- **USING THE OPTICAL DISK SYSTEM**

  The system is available for both staff and patrons. Security is not a problem as the whole configuration is in a secure area and at the current time is not networked outside of the NRL Library. The patron or the staff member sits down on a comfortable chair, identifies the needed reports on the STAR system at a terminal or PC, keys the accession numbers of the desired reports into the adjacent SUN work station and views and/or prints out any needed pages or reports in a pleasant atmosphere.

- **PLANNING FOR THE FUTURE**

  - Even though the system used to identify reports is separate from the Optical Disk System itself, plans are now being made to merge these systems. The indexing system would still be a separate entity but would have incorporated into it the "hooks" used to retrieve the report from the optical disk once the accession number is known.

  - When this collection consisting of 140,000 reports are scanned to disk, a second collection of 100,000 reports averaging 100 pages per item will also be put to disk. This will mean another 10 million pages of images stored on a separate autochanger to be daisy chained to the current one.

  - Scanning and printing out in color will be considered as the technology becomes more affordable.

- Plans are being made for the providing images over the campus network to the offices of the Naval Research Laboratory scientists. Future plans call for a study of the possibility of storing journal articles optically and networking their images to the scientists of the Laboratory as part of the Library's on-going plans to create an Electronic Library as a step in the direction of developing into a Library for the twenty first century.

- The NRL Library is talking to other libraries about sharing downloaded images of reports and journals by means of magnetic tape and/or 5 1/4 in. optical disks.

- Printing speeds will be improved. The Library has just installed a Laserjet HPIII si printer which when it is fully operational will print out images at a rate of 17 per minute.

- Scanners that not only scan in optical images but provide some optical character recognition (OCR) are being examined with a view to enlarging the retrieval data base. In the meantime a Kurtzweil OCR scanner is being utilized to place abstracts and parts of the main text of reports into the Cuadra Star data base.

- Once equipment like the Optical Disk System is in place, possible refinements become obvious and a wish list grows rather quickly - more work stations for patrons and staff, more information in the retrieval data base, faster scanners, faster printers, remote access and all the new products that this rapidly advancing technology is going to offer.

SUMMARY

The Library at the Naval Research Laboratory has designed and installed an optical disk system for the purpose of optical storage and retrieval of its large collection of reports. These reports, averaging 55 pages each, are scanned and their images captured on 12-inch optical disks stored in a jukebox holding fifty platters. Reports are identified in a separate database containing extensive indexing information pertaining to each report scanned. Once identified, a report can be retrieved from optical storage using the report's six digit accession number. The images of the retrieved report can be viewed on a screen or printed out. Forty thousand reports are currently stored optically with plans to complete this phase of the project by putting 100,000 additional reports to disk.

REFERENCES

1. Frank Moore Internal Revenue Service, Paper presented at the Association for Information and Image Management April 1988.

2. David A. Harvey and Bob Ryan, "Practically Paperless" *Byte*:185-190 (April 1991)

3. Thomas J Thiel, "Integrated CD-ROM and WORM Optical Disk Systems on the Navy's Paperless Ship" *CD-ROM Professional*: 17-26 (May 1992).

4. William L Hooton, Optical disk Storage System Project Report (Washington, D. C. National Archives and Records Administration March 1991).

5. Integrated Automation, Inc. "Facsimile Transmission on the Congressional Cable System". *Final Report for the Congressional Research Service* Library of Congress Washington D. C. (April 1987).

6. Ann R. Kann and Lynne K. Personius, "Update on digital Techniques". *Advances in Preservation and Access* Vol. 1 Meckler (1992)

7. "The Mercury Electronic Library and Library Information System II. The First Three years." Carnegie Mellon University Technical Report Series No. 6 (February 1992)

8. Masaaki Morita. Chairman, Sony Corporation of America. (Speech at the American Institute of Image Management April 1988).

9. Optical Disk Institute, P. O. Box 289, Newton, MA 02160

| Home | NRL | Suggest | Search |

webmaster@library.nrl.navy.mil
Updated: 27-NOV-95
Maintained by: Fred Rettenmaier